
aracne

Classificazione Decimale Dewey

— 004.6 (23.) INTERFACCIAMENTO E COMUNICAZIONI

Thema

— Soggetto: U. Informatica e Tecnologia dell'Informazione
— Qualificatori: 3MR. XXI secolo, 2000-2100

GERARDO **IOVANE**

DATA CENTER EUROPEI E NAZIONALI

HYPERSCALE DISTRIBUITI, MULTISCALA, CONFEDERALI

aracne



©

ISBN
979-12-218-3088-0

PRIMA EDIZIONE
ROMA 16 NOVEMBRE 2026

Indice

Prefazione 9

Introduzione 13

CAPITOLO 1 — IL DATA CENTER COME INFRASTRUTTURA FONDAMENTALE DELLA CIVILTÀ DIGITALE: DATO, CALCOLO, INTELLIGENZA ARTIFICIALE E LIMITI DELLA CONCENTRAZIONE 25

CAPITOLO 2 — STORIA DELLA CENTRALIZZAZIONE E DELLA DECENTRALIZZAZIONE DEL CALCOLO: DAL MAINFRAME ALLA CONFEDERAZIONE MULTISCALA 43

CAPITOLO 3 — PERCHÉ L'HYPERSCALE NON DEVE COINCIDERE CON UN EDIFICIO: SCALA FISICA, SCALA LOGICA E PROPRIETÀ EMERGENTI 59

CAPITOLO 4 — LA COOPERATIVE DIGITAL INFRASTRUCTURE COME FONDAMENTO TEORICO DELLA CONFEDERAZIONE 75

CAPITOLO 5 — FORMALIZZAZIONE DEL MULTISCALE CONFEDERATED HYPERSCALE: STRUTTURA, VETTORI DI RISORSA E CAPACITÀ DERIVATE 91

CAPITOLO 6 — C-ADCAS: IL CONFEDERATED AI DATA CENTER ARCHITECTURE STACK 107

CAPITOLO 7 — LEGGI DI SCALA COOPERATIVE: CAPACITÀ, DISPONIBILITÀ, DIVERSITÀ E SATURAZIONE DELLA COOPERAZIONE 123

CAPITOLO 8 — COMPUTATIONAL COUPLING THEORY: L'INDICE κW E LA CLASSIFICAZIONE DEI WORKLOAD 139

CAPITOLO 9 — EFFECTIVE CONFEDERATED COMPUTE: DAL CALCOLO INSTALLATO AL CALCOLO EFFETTIVAMENTE COOPERANTE 157

CAPITOLO 10 — IL MICRO DATA CENTER DA 100 kW: ARCHITETTURA, AUTONOMIA OPERATIVA ED ECONOMIA DELLA PROSSIMITÀ 173

CAPITOLO 11 — IL MINI DATA CENTER DA CIRCA 1 MW: AGGREGAZIONE TERRITORIALE, INTELLIGENZA ARTIFICIALE INDUSTRIALE E COMUNITÀ ENERGETICHE 189

CAPITOLO 12 — IL MEDIUM DATA CENTER DA DECINE DI MEGAWATT: CLOUD REGIONALE, FABBRICA DI MODELLI METROPOLITANA E SOVRANITÀ APPLICATA 207

CAPITOLO 13 — IL LARGE DATA CENTER FINO A CIRCA UN GIGAWATT: CAMPUS, MEGACLUSTER E ANCORAGGIO COMPUTAZIONALE DELLA CONFEDERAZIONE 223

CAPITOLO 14 — ARCHITETTURE IBRIDE OTTIMALI: CONFRONTO SISTEMATICO A PARITÀ DI POTENZA INSTALLATA 239

CAPITOLO 15 — ARCHITETTURA NAZIONALE ITALIANA A 6.012 NODI: TOPOLOGIA, GERARCHIA LOGICA E INSTRADAMENTO DEI CARICHI 257

CAPITOLO 16 — CAPACITÀ COMPUTAZIONALE DELLO SCENARIO ITALIA: POTENZA AGGREGATA, ANALISI DI SENSITIVITÀ E CALCOLO EFFETTIVAMENTE COOPERANTE 273

CAPITOLO 17 — CONFEDERAZIONE NAZIONALE CONTRO HYPERSCALE TRADIZIONALE: QUANDO VINCE IL MONOLITICO, QUANDO VINCE LA FEDERAZIONE, QUANDO VINCE L'IBRIDO 289

CAPITOLO 18 — NATIONAL AI COMPUTE BACKBONE: DORSALE NAZIONALE DEL CALCOLO, DISTRIBUZIONE DEI MODELLI E INFERENZA DI SISTEMA 307

CAPITOLO 19 — DAL SITE SELECTION AL NETWORK-OF-SITES SELECTION: LA LOCALIZZAZIONE COME PROBLEMA DI RETE 323

CAPITOLO 20 — MATEMATICA DELLA LOCALIZZAZIONE: MODELLI DI FACILITY LOCATION, PARTIZIONAMENTO DEL GRAFO E OTTIMIZZAZIONE MULTI-OBIETTIVO 339

CAPITOLO 21 — LE AREE INTERNE COME INFRASTRUTTURA DIGITALE PRODUTTIVA: CALCOLO, LAVORO E COESIONE TERRITORIALE 355

CAPITOLO 22 — L'ENERGIA DEL CONFEDERATED HYPERSCALE: PROFILO DI CARICO, DIVERSITÀ E SERVIZI ALLA RETE 371

CAPITOLO 23 — FONTI RINNOVABILI DISTRIBUITE: FOTOVOLTAICO, EOLICO, IDROELETTRICO E OTTIMIZZAZIONE DI PORTAFOGLIO 389

CAPITOLO 24 — SISTEMI DI ACCUMULO ELETTROCHIMICO: DAL GRUPPO DI CONTINUITÀ ALL'ACCUMULO DI FEDERAZIONE 407

CAPITOLO 25 — COMUNITÀ ENERGETICHE E COMUNITÀ DI DATI: IL PROSUMER DIGITALE E LA CONTABILITÀ DEL CONFERIMENTO 425

CAPITOLO 26 — MIGRAZIONE DEI CARICHI CONSAPEVOLE DEL CARBONIO: SCHEDULING TEMPORALE, GEOGRAFICO E LIMITI DEL METODO 443

CAPITOLO 27 — TERMODINAMICA DEL CALCOLO DISTRIBUITO: DENSITÀ, TECNOLOGIE DI RAFFREDDAMENTO E VINCOLI DI SCALA 461

CAPITOLO 28 — RECUPERO DEL CALORE E TELERISCALDAMENTO: QUALITÀ TERMICA, SCALA DELL'UTENZA E MODELLI CONTRATTUALI 477

CAPITOLO 29 — RAFFREDDAMENTO CONSAPEVOLE DEL CLIMA E ORCHESTRAZIONE TERMICA: METRICHE, COMPROMESSI E LIMITI 493

CAPITOLO 30 — LA RETE COME COMPONENTE COMPUTAZIONALE: TOPOLOGIA, DIMENSIONAMENTO E COSTO STRUTTURALE DELLA DISTRIBUZIONE 511

CAPITOLO 31 — TECNOLOGIE DI TRASPORTO E FABRIC CONFEDERALE: OTTICA, LATENZA DETERMINISTICA E SEGMENTAZIONE 527

CAPITOLO 32 — ORCHESTRAZIONE CONGIUNTA DI RETE E CALCOLO: CO-SCHEDULING, PRENOTAZIONE E DEGRADO CONTROLLATO 543

CAPITOLO 33 — CALCOLO CONFEDERATO: ETEROGENEITÀ, ASTRAZIONE DELLA RISORSA E POOL CONDIVISO 561

CAPITOLO 34 — ARCHIVIAZIONE CONFEDERATA: REPLICA, CODIFICA, COERENZA E DURABILITÀ DISTRIBUITA 577

CAPITOLO 35 — IL DATO SOVRANO: CLASSIFICAZIONE, PREDICATI DI AMMISSIBILITÀ E PREZZO DELLA SOVRANITÀ 593

CAPITOLO 36 — INTELLIGENZA ARTIFICIALE NEL CONFEDERATED HYPERSCALE: CICLO DI VITA DEL MODELLO E DISTRIBUZIONE FRA LE SCALE 609

CAPITOLO 37 — L'INTELLIGENZA ARTIFICIALE COME STRUMENTO DI GOVERNO DELL'INFRASTRUTTURA: PREVISIONE, ALLOCAZIONE E DECISIONE INCERTA 625

CAPITOLO 38 — AGENTI E SISTEMI MULTI-AGENTE PER L'ORCHESTRAZIONE CONFEDERALE: NEGOZIAZIONE, FIDUCIA E RESPONSABILITÀ 641

CAPITOLO 39 — INTELLIGENZA ARTIFICIALE SOVRANA E MODELLI TERRITORIALI: CATENA DI DIPENDENZA, STRATEGIE DI AUTONOMIA E PLURALISMO 657

CAPITOLO 40 — IL SISTEMA OPERATIVO DELLA CONFEDERAZIONE: PIANO DI CONTROLLO, IDENTITÀ E INTEROPERABILITÀ 675

CAPITOLO 41 — SCHEDULING E ALLOCAZIONE MULTI-OBIETTIVO: CLASSI DI SERVIZIO, EURISTICHE PRATICABILI ED EQUITÀ 691

CAPITOLO 42 — INTEROPERABILITÀ, STANDARD E INTERFACCE DELLA CONFEDERAZIONE: VOCABOLARI, CONFORMITÀ E GOVERNO 707

CAPITOLO 43 — SICUREZZA DELL'INFRASTRUTTURA CONFEDERALE: MODELLO DI MINACCIA, FIDUCIA ZERO E PARTECIPANTE OSTILE 723

CAPITOLO 44 — RESILIENZA E CONTINUITÀ: GUASTI CORRELATI, RIDONDANZA COOPERATIVA E DEGRADO CONTROLLATO 739

CAPITOLO 45 — VERIFICA SPERIMENTALE DELLA RESILIENZA: INGEGNERIA DEL CAOS, RAGGIO DI IMPATTO E GOVERNO DEGLI ESPERIMENTI 755

CAPITOLO 46 — GESTIONE DEGLI INCIDENTI E APPRENDIMENTO ORGANIZZATIVO: DIAGNOSI, COMANDO E FATTORI SISTEMICI 771

CAPITOLO 47 — CALCOLO CONFIDENZIALE E PROTEZIONE DEL DATO IN USO: AMBIENTI FIDATI, CRITTOGRAFIA AVANZATA E SCELTA DELLA TECNICA 787

CAPITOLO 48 — RESILIENZA QUANTISTICA: MINACCIA, MIGRAZIONE POST-QUANTISTICA E ARCHITETTURA DELLE CHIAVI 803

CAPITOLO 49 — CRESCITA DELLA DOMANDA E LIMITI MATERIALI: RIMBALZO, SUFFICIENZA E ALLOCAZIONE DELLA CAPACITÀ 819

CAPITOLO 50 — PIANIFICAZIONE E INVESTIMENTO: DIMENSIONAMENTO, VALUTAZIONE ECONOMICA E SEQUENZA DI REALIZZAZIONE 835

CAPITOLO 51 — GOVERNANCE DELLA CONFEDERAZIONE: SOGGETTI, REGOLE E PROCESSI DECISIONALI 851

CAPITOLO 52 — ECONOMIA DELLA FEDERAZIONE: CREDITI DI RISORSA, TARIFFE E CONDIZIONI DI PERSISTENZA 867

CAPITOLO 53 — IL QUADRO REGOLATORIO: DISCIPLINE APPLICABILI, LACUNE E ADATTAMENTI NECESSARI 883

CAPITOLO 54 — L'ASSETTO ISTITUZIONALE: COMPETENZE, COORDINAMENTO MULTILIVELLO E PERCORSO PRATICABILE 899

CAPITOLO 55 — LA DIMENSIONE EUROPEA: CHE COSA FEDERARE A SCALA CONTINENTALE E CON QUALE GOVERNO 915

CAPITOLO 56 — SCENARI: TRAIETTORIE, BIFORCAZIONI E INDICATORI DI PERCORSO 931

CAPITOLO 57 — IL PROSUMER DIGITALE TERRITORIALE: CONTRIBUTO, CREDITI E CATTURA LOCALE DEL VALORE 947

CAPITOLO 58 — LA CONFEDERAZIONE COME POLITICA INDUSTRIALE: SOVRANITÀ, COESIONE E AUTONOMIA STRATEGICA 965

CAPITOLO 59 — ARCHITETTURA EUROPEA DELLE CONFEDERAZIONI NAZIONALI: INTEROPERABILITÀ SENZA PERDITA DI SOVRANITÀ 983

CAPITOLO 60 — ERASMUS: LA CONFEDERAZIONE EUROPEA DEL CALCOLO 1003

CAPITOLO 61 — BENCHMARK GLOBALE: HYPERSCALE TRADIZIONALI, CONFEDERAZIONI NAZIONALI ED ERASMUS A CONFRONTO 1021

CAPITOLO 62 — IL CONFEDERATED HYPERSCALE SCORE: *CHSCompute*, *CHSSystem*, *CHSAI* E PUNTEGGIO COMPLESSIVO 1039

CAPITOLO 63 — IL GRANDE ESPERIMENTO COMPARATIVO: SEI SCENARI A PARITÀ DI CAPACITÀ E TRENTATRÉ METRICHE 1055

CAPITOLO 64 — LA CONFEDERATED HYPERSCALE DOMINANCE MAP: QUANDO CIASCUNA SCALA DOMINA 1073

CAPITOLO 65 — IL TEOREMA DELLA SCALA OTTIMALE: ENUNCIATO, CONDIZIONI DI VALIDITÀ E CONSEGUENZE PROGETTUALI 1089

CONCLUSIONI 1105

APPENDICE A — MODELLO MATEMATICO FORMALE DELLA CONFEDERAZIONE 1111

APPENDICE B — AFFIDABILITÀ E DISPONIBILITÀ EMERGENTE 1117

APPENDICE C — MODELLI DELL'ACCOPPIAMENTO COMPUTAZIONALE 1121

APPENDICE D — MODELLI DEL CALCOLO EFFETTIVAMENTE COOPERANTE 1125

APPENDICE E — MODELLI ENERGETICI 1131

APPENDICE F — MODELLI TERMICI 1135

APPENDICE G — MODELLI DI RETE 1139

APPENDICE H — MODELLI DEI CARICHI DI INTELLIGENZA ARTIFICIALE 1143

APPENDICE I — MODELLI DI CAPITALE, ESERCIZIO E COSTO TOTALE 1149

APPENDICE J — IL QUADRO REGOLATORIO APPLICABILE 1153

APPENDICE K — METODOLOGIA DI SIMULAZIONE E RIPRODUCIBILITÀ 1159

APPENDICE L — ARCHITETTURE DI RIFERIMENTO DA 100 kW A 1 GW 1165

APPENDICE M — ATLANTE ITALIANO: STRUTTURA, VARIABILI E FONTI 1171

APPENDICE N — ATLANTE EUROPEO: STRUTTURA, VARIABILI E FONTI 1177

Prefazione

Il volume che precede questo, Data Center nell'Era dell'Intelligenza Artificiale, si chiudeva con una constatazione che allora sembrava soltanto un limite del metodo e che con il tempo si è rivelata l'apertura di un problema nuovo. Dopo aver descritto il data center nella sua materialità, dalla selezione del sito alla struttura portante, dalla distribuzione elettrica alla catena del freddo, dai cluster di acceleratori alle reti di interconnessione, dopo averne discusso l'economia, la sostenibilità e la regolazione, restava una domanda che l'impostazione di quel lavoro non poteva accogliere: fino a che punto ha senso continuare a studiare il data center come un oggetto, quando le grandezze che ne determinano il comportamento sono ormai grandezze di sistema. La potenza disponibile non è una proprietà dell'impianto ma del nodo di rete elettrica a cui l'impianto si allaccia. La latenza percepita dall'utente non dipende dall'architettura interna del sito ma dalla sua posizione rispetto alla domanda. La sostenibilità non si esaurisce nel rapporto fra energia assorbita ed energia utile, perché l'acqua consumata, il suolo occupato, il calore disperso e il carbonio incorporato sono quantità che si misurano sul territorio e non sul perimetro della recinzione. La disponibilità del servizio, infine, è da tempo una proprietà emergente di insiemi di siti, non del singolo sito e nessun progettista serio la calcola più a partire dalla ridondanza locale soltanto.

Da queste constatazioni nasce il presente volume. La sua ipotesi di lavoro è semplice da enunciare e impegnativa da sostenere: l'unità di analisi rilevante per l'infrastruttura computazionale contemporanea non è più il data center, ma la confederazione di data center di scala diversa che coopera su un territorio. Non si tratta di un cambio di prospettiva puramente descrittivo. Se l'unità di analisi cambia, cambiano le grandezze che la misurano, cambiano i criteri di progetto, cambiano i modelli di costo e cambia il modo in cui un sistema-paese può pensare la propria capacità di calcolo. Un volume dedicato a questo spostamento non è dunque un'appendice del precedente, ma la sua evoluzione necessaria, allorquando si desideri avere una visione che si armonizzi con il territorio.

L'occasione immediata è stata offerta dal lavoro di ricerca svolto negli ultimi anni sulle infrastrutture digitali cooperative, sui micro data center federati e sui meccanismi che ne rendono possibile la governance distribuita in condizioni di fiducia limitata. Quel lavoro, nato in un contesto apparentemente laterale, ha mostrato che i problemi teorici sollevati dalla federazione di nodi piccoli non sono diversi, nella sostanza, da quelli sollevati dalla federazione di nodi grandi: si tratta sempre di stabilire quando la somma delle capacità è operativamente significativa, a quale costo la cooperazione può avvenire e quali proprietà emergono dall'insieme che nessuna delle parti possiede. Portare quelle domande alla scala nazionale e continentale è stato il passo successivo e questo libro ne è il risultato.

Occorre dire subito quale posizione l'opera assume nel dibattito pubblico sulla concentrazione del calcolo, perché quel dibattito è oggi fortemente polarizzato e la polarizzazione produce cattiva ingegneria. Da un lato si sostiene che soltanto la mega-facility

consenta di addestrare i modelli che definiscono la frontiera tecnologica e che ogni discorso sulla distribuzione sia una forma di consolazione per chi non dispone del capitale necessario. Dall'altro si sostiene che la concentrazione sia un errore in sé, ambientalmente insostenibile, socialmente inaccettabile e tecnicamente superabile da una rete sufficientemente densa di nodi piccoli. Nessuna delle due posizioni resiste a un esame quantitativo. La prima ignora che una quota larghissima della domanda computazionale reale è debolmente accoppiata, sensibile alla latenza e vincolata alla località del dato e che per quella quota la mega-facility è una soluzione costosa e inefficiente. La seconda ignora che esistono classi di carichi la cui esecuzione richiede una prossimità fisica fra acceleratori che nessuna rete geografica, per quanto capace, è in grado di surrogare. La tesi di questo libro sta altrove ed è più esigente di entrambe: ogni scala infrastrutturale possiede un proprio dominio di ottimalità, quel dominio è determinabile a partire dalle caratteristiche del carico e del contesto e il compito dell'ingegneria delle infrastrutture non è scegliere una scala ma orchestrarle tutte dentro un'unica architettura.

Questa posizione ha una conseguenza che vale la pena anticipare. Il volume non contrappone la confederazione all'hyperscale: la confederazione, come qui viene definita, contiene l'hyperscale come proprio livello superiore. Chi cercasse in queste pagine un argomento contro i grandi campus di calcolo resterà deluso, così come resterà deluso chi cercasse la dimostrazione che la capillarità territoriale sia di per sé un valore. La prima tesi forte del libro è che l'aggettivo hyperscale può qualificare un sistema distribuito e non soltanto un edificio; la seconda è che la geografia, l'energia, la rete e il territorio vanno trattati come componenti dell'architettura computazionale, non come vincoli esterni che il progetto subisce.

Il libro si rivolge a quattro tipi di lettore. Il ricercatore vi troverà un apparato formale che rende quantitative nozioni finora trattate in modo discorsivo, in particolare quelle di accoppiamento computazionale, di capacità effettivamente cooperante e di scala sufficiente. L'ingegnere e il progettista vi troveranno criteri di dimensionamento, confronti sistematici fra classi di nodo e la discussione esplicita dei punti in cui la distribuzione geografica smette di essere conveniente. L'economista e il decisore pubblico vi troveranno l'analisi dei costi, dei modelli di proprietà e delle implicazioni di politica industriale, insieme a un quadro delle condizioni sotto cui la capacità di calcolo diventa un attributo di sovranità e non soltanto una voce di bilancio. L'amministratore locale, infine, vi troverà la ragione per cui un nodo di calcolo installato in un territorio marginale può produrre valore misurabile e insieme le condizioni perché ciò accada realmente invece di restare una promessa.

Un avvertimento metodologico è dovuto fin da queste prime righe. Il volume fa largo uso di scenari numerici nazionali ed europei, costruiti per rendere confrontabili architetture alternative a parità di capacità installata. Quegli scenari non sono previsioni, non sono piani di investimento e non intendono suggerire che una certa configurazione debba essere realizzata. Sono strumenti di ragionamento e come tali vanno letti: la loro funzione è mostrare come si comportano grandezze diverse quando la stessa capacità viene organizzata

in modi diversi, non indicare quante strutture un paese debba costruire. Chi estraesse da queste pagine cifre da usare come obiettivi di programmazione ne farebbe un uso improprio, contro il quale il testo mette in guardia ripetutamente.

Resta il debito di riconoscenza. Il lavoro che ha condotto a questo volume si è nutrito del confronto quotidiano con colleghi e collaboratori del team coordinato da chi scrive, che hanno accettato di discuterne le ipotesi quando erano ancora fragili e che ne hanno segnalato le debolezze con la franchezza che soltanto la stima reciproca rende possibile. Un ringraziamento particolare va all'ambiente universitario nel quale la ricerca si è svolta, che ha garantito la libertà di percorrere una strada non ancora consolidata e agli studenti dei corsi di infrastrutture digitali, la cui insistenza nel chiedere il perché di ogni soglia e di ogni coefficiente ha migliorato la formulazione di molte delle idee che qui si trovano. Un ringraziamento va inoltre ai tecnici, ai progettisti e agli operatori che hanno accettato di descrivere problemi reali di esercizio: il contatto con la pratica impedisce alla teoria di diventare elegante e inutile. Gli errori, le semplificazioni eccessive e le posizioni discutibili che il lettore troverà in queste pagine restano naturalmente responsabilità esclusiva di chi scrive.

Un'ultima osservazione riguarda il momento in cui il libro viene pubblicato. L'infrastruttura del calcolo attraversa una fase di espansione che non ha precedenti per intensità di capitale e per rapidità di obsolescenza; ogni testo che pretenda di descriverla corre il rischio di invecchiare prima di arrivare al lettore. Per ridurre quel rischio il volume privilegia deliberatamente le grandezze strutturali rispetto ai dati congiunturali, i rapporti fra quantità rispetto ai valori assoluti, i criteri di decisione rispetto alle soluzioni tecnologiche del momento. Le generazioni di acceleratori cambieranno, i costi per unità di capacità caleranno, le densità di potenza per armadio cresceranno ancora; ciò che non cambierà, almeno finché la velocità della luce resta quella che è, è il fatto che la comunicazione costa tempo e che il tempo speso a comunicare non è tempo speso a calcolare. Su questa asimmetria elementare poggia l'intero edificio argomentativo del volume.

Nella realizzazione di quest'opera sono stati utilizzati strumenti di intelligenza artificiale a supporto di traduzione, ricerca documentale e rielaborazione stilistica del testo. L'autore ha effettuato personalmente la verifica, il controllo dei contenuti e l'approvazione finale del testo.

Introduzione

I.1 Il contesto: la trasformazione strutturale dell'infrastruttura digitale globale

L'infrastruttura digitale mondiale sta attraversando la trasformazione più profonda dalla nascita del cloud e la natura di questa trasformazione è spesso descritta in modo impreciso. Non si tratta soltanto di una crescita quantitativa della capacità installata, per quanto tale crescita sia imponente e senza precedenti nella storia dell'infrastruttura tecnica. Si tratta di un mutamento qualitativo della domanda, che modifica simultaneamente la densità di potenza dei sistemi, la struttura dei costi, i vincoli di localizzazione e il rapporto fra infrastruttura digitale e sistema elettrico. Ciascuno di questi mutamenti, preso isolatamente, sarebbe gestibile con gli strumenti consolidati della disciplina. La loro simultaneità è ciò che produce la discontinuità.

Il primo elemento della trasformazione è la mutata composizione del carico. Per oltre un decennio la domanda dominante è stata quella dei servizi cloud generalisti, caratterizzata da un'intensità computazionale moderata, da una tolleranza relativamente ampia sulla latenza e da una distribuzione temporale prevedibile. L'ingresso su scala industriale dei carichi di apprendimento automatico ha introdotto una domanda di natura diversa, concentrata nel tempo, estremamente intensiva nell'uso degli acceleratori e caratterizzata da requisiti di comunicazione interna che non hanno equivalenti nel calcolo tradizionale. La conseguenza immediata è stata l'aumento della densità di potenza per armadio, cresciuta di un ordine di grandezza nell'arco di pochi anni, con effetti a cascata sull'architettura elettrica, sulla catena del freddo e sulla progettazione strutturale dei locali tecnici.

Il secondo elemento è l'emergere del vincolo energetico come vincolo primario. Per lungo tempo la disponibilità di potenza elettrica è stata un parametro di progetto fra gli altri, risolvibile con un allacciamento adeguato. Oggi la capacità di connessione alla rete di trasmissione è divenuta il fattore che determina dove un impianto può sorgere e quando può entrare in esercizio, con tempi di attesa che in diverse aree europee superano la durata stessa della costruzione dell'impianto. Questo rovesciamento ha una portata che va oltre la logistica: significa che l'infrastruttura digitale è entrata in competizione diretta con gli altri usi dell'elettricità e che il suo sviluppo dipende ormai da decisioni di pianificazione che non appartengono al settore digitale.

Il terzo elemento è la comparsa dei vincoli ambientali e territoriali come condizioni vincolanti e non come oneri accessori. Il consumo idrico dei sistemi di raffreddamento, il consumo di suolo, l'impatto sul paesaggio, il contributo alle emissioni climalteranti e la contesa con le comunità locali sui procedimenti autorizzativi hanno smesso di essere questioni marginali della fase realizzativa e sono diventati fattori che determinano la fattibilità stessa dei progetti di grande taglia. Nelle aree a maggiore concentrazione di infrastruttura la resistenza sociale all'insediamento di nuovi campus ha raggiunto intensità

tali da produrre moratorie amministrative, con effetti diretti sulla geografia degli investimenti.

Il quarto elemento è la trasformazione della capacità di calcolo in attributo di sovranità. Quando il calcolo diventa il fattore produttivo su cui poggiano la ricerca scientifica, la difesa, la sanità, l'amministrazione pubblica e una quota crescente dell'attività economica, la sua collocazione geografica e giuridica smette di essere indifferente. Le politiche di localizzazione dei dati, i requisiti di certificazione dei fornitori e le restrizioni all'esportazione di componenti avanzate hanno prodotto una frammentazione del mercato globale che nessun operatore può ignorare. La capacità computazionale disponibile entro una giurisdizione e sotto controllo giuridico di quella giurisdizione è diventata una grandezza di interesse pubblico.

Il quinto elemento, meno discusso ma non meno rilevante, riguarda la struttura della domanda territoriale. La crescita dei servizi digitali non è distribuita in modo uniforme, ma neppure è concentrata dove si concentra l'infrastruttura. Le applicazioni che richiedono tempi di risposta brevi e continuità di servizio nascono dove vivono le persone e dove operano le imprese, quindi in modo capillare sul territorio, mentre l'infrastruttura si concentra in un numero ristretto di poli scelti per ragioni di costo dell'energia, di disponibilità di connettività e di regime fiscale. Questa divaricazione fra geografia della domanda e geografia dell'offerta produce un costo, che si manifesta in latenza, in traffico di transito, in dipendenza da infrastrutture di trasporto e in fragilità sistemica.

L'insieme di questi cinque elementi definisce il contesto entro cui si colloca l'argomento di questo volume. La domanda computazionale cresce più rapidamente della capacità di assorbirla con il paradigma della concentrazione, i vincoli materiali che limitano la concentrazione non sono superabili con l'innovazione tecnologica nel breve periodo, la domanda reale è geograficamente distribuita e la dimensione istituzionale della capacità di calcolo impone criteri di localizzazione che l'ottimizzazione puramente economica non contempla. In queste condizioni la questione non è se l'infrastruttura si distribuirà, perché la distribuzione è già in corso sotto la spinta dei vincoli, ma se tale distribuzione avverrà in modo governato e progettato oppure in modo incrementale e incoerente. Il volume si propone di fornire gli strumenti concettuali del primo esito.

I.2 Dal data center alla confederazione: perché il paradigma della mega-facility non esaurisce il problema

Il paradigma dominante degli ultimi quindici anni ha identificato il progresso dell'infrastruttura computazionale con la crescita dimensionale della singola struttura. Tale identificazione non è stata arbitraria: essa riflette economie di scala reali e documentate, che riguardano il costo unitario dell'energia, l'efficienza degli impianti di raffreddamento, la produttività del personale tecnico, il potere contrattuale nell'approvvigionamento dei componenti e la possibilità di co-progettare hardware e software su volumi che giustificano lo sviluppo di silicio dedicato. Chi sostiene che la concentrazione sia stata un errore ignora

la solidità di queste economie. Il punto non è negarle, ma riconoscere che ogni curva di rendimento di scala possiede un tratto terminale in cui i benefici marginali si esauriscono mentre i costi marginali continuano a crescere e che l'infrastruttura digitale ha raggiunto quel tratto.

I segnali dell'esaurimento sono riconoscibili. Il costo marginale dell'allacciamento elettrico cresce più che proporzionalmente con la taglia, perché oltre una certa soglia non esiste un nodo di rete capace di assorbire il carico senza interventi di rinforzo che hanno tempi e costi propri. Il costo marginale del raffreddamento cresce con la densità, perché il passaggio dall'aria al liquido comporta una discontinuità impiantistica e organizzativa. Il costo marginale del consenso territoriale cresce con la concentrazione, perché l'impatto percepito da una comunità non è funzione lineare della potenza installata. Il costo marginale del rischio cresce con la concentrazione della capacità, perché la perdita di un sito che ospita una quota significativa della capacità nazionale produce un danno sistemico non compensabile. Nessuno di questi effetti è nuovo; ciò che è nuovo è che si manifestano tutti contemporaneamente e nella stessa fascia dimensionale.

A questa constatazione si affianca un argomento di natura diversa, che riguarda non i costi ma l'idoneità. Una parte della domanda computazionale è intrinsecamente non delocalizzabile: i dati clinici che non possono lasciare la struttura sanitaria, i processi industriali che richiedono tempi di risposta dell'ordine del millisecondo, i sistemi di sorveglianza urbana che generano flussi la cui trasmissione integrale a un sito remoto sarebbe più costosa della loro elaborazione locale, i servizi pubblici che devono continuare a funzionare quando la connettività geografica si interrompe. Per questa parte della domanda il campus remoto non è una soluzione costosa, è una soluzione inapplicabile. La distribuzione dell'infrastruttura non è dunque una scelta, ma la risposta obbligata a una classe di requisiti.

Il paradigma dell'edge computing ha riconosciuto questa necessità, ma l'ha affrontata in modo parziale. L'edge è stato concepito come estensione periferica del cloud, come strato sottile di prossimità che filtra, aggrega e preelabora il dato prima di inviarlo al centro. Tale concezione conserva intatta la gerarchia: il centro calcola, la periferia prepara. Essa non risponde alla domanda che qui interessa, che riguarda la possibilità che nodi di scala diversa cooperino su un piano di sostanziale parità funzionale, contribuendo ciascuno a un'unica capacità di sistema anziché occupare posizioni fisse in una piramide.

La nozione di confederazione nasce da questo scarto. Una confederazione di data center non è una gerarchia con un vertice e una base, ma una rete nella quale il carico può salire verso nodi di scala maggiore, scendere verso nodi di scala minore oppure migrare lateralmente fra nodi di scala equivalente, in funzione dei requisiti del carico stesso e dello stato del sistema. Il nodo di grande taglia resta indispensabile per ciò che soltanto esso può fare, ma smette di essere il centro logico dell'architettura per diventarne uno dei livelli. La capacità che il sistema offre non è la capacità del nodo maggiore né la somma indiscriminata delle capacità

dei nodi, ma una grandezza che dipende da quanto efficacemente i nodi riescono a cooperare su un dato carico.

Da questo spostamento discende la necessità di un apparato concettuale nuovo, perché le grandezze consolidate non sono in grado di descrivere il sistema. L'efficienza energetica dell'infrastruttura, misurata come rapporto fra energia complessivamente assorbita ed energia utile, è definita sul perimetro di un impianto e perde significato quando il perimetro è distribuito su un territorio. La disponibilità del servizio, calcolata a partire dalla ridondanza degli impianti, non cattura la ridondanza che deriva dalla molteplicità dei siti. La capacità di calcolo, espressa come numero di operazioni al secondo installate, non dice nulla sulla frazione di quelle operazioni che può concorrere allo stesso compito. Occorrono grandezze che misurino la cooperazione, non l'installazione ed è questa la ragione per cui il volume dedica la propria seconda Parte alla costruzione di tale apparato prima di affrontare qualunque questione ingegneristica.

I.3 La tesi del volume, i suoi corollari e i suoi limiti di validità dichiarati

La tesi centrale del volume può essere enunciata in una proposizione: l'hyperscale non è necessariamente una proprietà dimensionale del singolo data center, ma può costituire una proprietà emergente della confederazione. L'aggettivo che per un decennio ha qualificato edifici e campus qualifica in questa prospettiva un comportamento di sistema, riconoscibile quando un insieme di nodi coopera in modo tale da offrire capacità, disponibilità e flessibilità che nessuno dei nodi possiede singolarmente. L'enunciato non è metaforico. Esso è verificabile e il volume indica le condizioni sotto cui risulta vero e quelle sotto cui risulta falso.

La tesi complementare riguarda lo statuto del contesto fisico e istituzionale. La geografia, l'energia, la rete e il territorio non sono vincoli esterni che il progetto subisce, ma componenti dell'architettura computazionale, al pari della memoria e del fabric di interconnessione. Questa affermazione ha conseguenze operative precise: la distanza fra due nodi è un parametro architetturale nello stesso senso in cui lo è la latenza di accesso alla memoria; il profilo orario del prezzo dell'energia è un parametro di scheduling nello stesso senso in cui lo è la priorità del processo; la giurisdizione entro cui un dato risiede è un vincolo di allocazione nello stesso senso in cui lo è la capacità residua di un nodo. Trattare questi elementi come contesto anziché come componenti produce architetture che funzionano sulla carta e falliscono nell'esercizio.

Dalla tesi discende un principio strategico che merita di essere sottolineato per evitare fraintendimenti ricorrenti: il paradigma distribuito non elimina il data center di grande taglia, lo incorpora. La confederazione contiene i nodi da centinaia di megawatt come propri elementi costitutivi e ne riconosce l'insostituibilità per la classe di carichi che richiede prossimità fisica fra un numero elevato di acceleratori. Chi legga il volume come un manifesto contro la concentrazione ne fraintende l'argomento. La posizione sostenuta è che

la concentrazione sia necessaria per una parte della domanda e inefficiente per la parte restante; il compito dell'architettura è riconoscere il confine fra le due.

Il principio operativo che regge l'intera costruzione è quello di minima scala sufficiente: ogni carico deve essere eseguito alla minima scala infrastrutturale capace di soddisfarne simultaneamente i requisiti di calcolo, memoria, banda, latenza, resilienza, sostenibilità e sovranità. Il principio è normativo, non descrittivo: esso indica come dovrebbe essere allocato il carico, non come viene allocato nella pratica corrente. La sua applicazione richiede che i requisiti siano esprimibili in forma quantitativa e che le capacità delle classi di nodo siano note, condizioni che il volume si preoccupa di rendere verificabili.

I corollari della tesi sono tre. Il primo è che non esiste una scala universalmente ottimale, ma una scala ottimale in funzione del carico e del contesto, determinabile con un procedimento esplicito. Il secondo è che il vantaggio della federazione non è uniforme rispetto ai carichi, ma cresce al diminuire dell'accoppiamento e al crescere della sensibilità alla latenza, ai vincoli di località del dato e al peso dei fattori energetici e territoriali. Il terzo è che la cooperazione ha un costo, che si manifesta in complessità di orchestrazione, in traffico geografico e in oneri di governance; tale costo definisce una soglia oltre la quale aggiungere nodi alla confederazione non produce più valore.

Vanno infine dichiarati i limiti di validità. Il volume argomenta su un piano teorico e modellistico, non sperimentale: le architetture confrontate sono costruzioni analitiche valutate attraverso modelli, non sistemi realizzati e misurati. I parametri dei modelli derivano dalla letteratura tecnica, dalla documentazione pubblica degli operatori e dall'esperienza progettuale, con un'incertezza che in alcuni casi è considerevole e che viene dichiarata caso per caso. Le conclusioni relative ai costi dipendono da ipotesi sui prezzi dell'energia, del capitale e dell'hardware che sono valide nell'orizzonte in cui il volume è scritto e che potrebbero non esserlo oltre. Le conclusioni relative alla sostenibilità dipendono dalla qualità dei dati disponibili sulle emissioni incorporate, che è tuttora insoddisfacente. Il volume, infine, adotta una prospettiva europea; le sue conclusioni istituzionali e regolatorie non sono trasferibili senza adattamenti ad altri ordinamenti. Dichiarare questi limiti non indebolisce l'argomento: ne definisce il dominio di applicabilità, che è ciò che distingue un'ipotesi scientifica da una posizione ideologica.

I.4 Le grandezze fondamentali del Distributed Hyperscale: accoppiamento, calcolo cooperante, scala sufficiente e punteggio confederale

La costruzione teorica del volume poggia su un piccolo numero di grandezze, introdotte qui nella loro forma essenziale e riprese in modo sistematico nella seconda Parte. La prima e più importante è l'indice di accoppiamento computazionale, indicato con $\kappa_W \in [0,1]$ e riferito a un carico di lavoro W. Esso misura il grado in cui l'esecuzione del carico richiede comunicazione e sincronizzazione fra le unità di calcolo che vi partecipano. Un valore prossimo all'unità caratterizza i carichi fortemente accoppiati, dove ogni passo di elaborazione è preceduto e seguito da uno scambio di informazione fra tutte le unità

coinvolte, come accade nell'addestramento dei modelli di frontiera e nelle simulazioni scientifiche a grana fine. Un valore prossimo a zero caratterizza i carichi in cui le unità operano in modo sostanzialmente indipendente, come l'inferenza su richieste separate, la distribuzione di contenuti o l'elaborazione di flussi provenienti da sensori. L'indice non è una etichetta qualitativa: esso si stima a partire da grandezze osservabili, prima fra tutte il rapporto fra tempo speso in comunicazione e tempo speso in calcolo.

Dall'accoppiamento discende il criterio che governa la distribuibilità geografica. Poiché la latenza fra nodi distanti supera di tre ordini di grandezza la latenza interna a un nodo, un carico con κ_W elevato subisce nella distribuzione una penalizzazione che può annullare per intero il vantaggio della capacità aggiuntiva, mentre un carico con κ_W basso la subisce in misura trascurabile. Questa asimmetria, che deriva da una proprietà fisica e non da una limitazione tecnologica superabile, è il fondamento dell'intero volume e la ragione per cui la confederazione non può essere presentata come soluzione universale.

La seconda grandezza traduce in termini quantitativi la distinzione fra capacità installata e capacità utilizzabile. Il calcolo effettivamente cooperante si esprime come

$$\mathrm{ECC} = C_{\mathrm{raw}} \cdot \eta_{\mathrm{HW}} \cdot \eta_{\mathrm{util}} \cdot \eta_{\mathrm{net}} \cdot \eta_{\mathrm{place}},$$

dove C_{raw} è la capacità teorica installata, η_{HW} l'efficienza dell'hardware rispetto al suo valore di picco, η_{util} l'utilizzazione effettiva, η_{net} il fattore che riassume la penalizzazione dovuta alla rete e η_{place} la qualità dell'allocazione operata dall'orchestratore. La struttura moltiplicativa è significativa: poiché ciascun fattore è minore dell'unità, un sistema può disporre di una capacità nominale imponente e di una capacità cooperante modesta e il divario fra le due cresce con l'accoppiamento del carico. Accanto a questa grandezza si colloca la capacità di servizio distribuito, $Q_{\mathrm{tot}} = \sum_i Q_i$, che misura invece la capacità complessiva di servire carichi indipendenti e per la quale l'aggregazione è legittima.

La terza grandezza formalizza il principio operativo del volume. La scala sufficiente per un carico è definita come

$$s^{(W)} = \min_{s} s : C_s \geq C_W, ; M_s \geq M_W, ; B_s \geq B_W, ; L_s \leq L_W, ; R_s \geq R_W, ; S_s \geq S_W,$$

dove i simboli indicano rispettivamente i requisiti di calcolo, memoria, banda, latenza, resilienza e sovranità, confrontati con le capacità corrispondenti della classe di nodo s. La formulazione è volutamente elementare, perché la sua difficoltà non sta nella matematica ma nella determinazione dei requisiti e delle capacità. In forma più generale la scala ottimale si scrive come

$$\mathrm{Scale}^{*}(W) = f(\kappa_W, L_W, B_W, C_W, M_W, D_W, E, S_W, R_W, \mathrm{Cost}),$$

espressione che riassume la dipendenza dal carico e dal contesto.

La quarta grandezza riguarda la disponibilità. Per nodi con guasti indipendenti la disponibilità di sistema si scrive come

$$A_{\text{sys}} = 1 - \prod_i (1 - A_i),$$

espressione che cresce rapidamente con il numero dei nodi e che spiega perché la molteplicità dei siti sostituisca efficacemente la ridondanza degli impianti. L'ipotesi di indipendenza, tuttavia, è la più fragile dell'intera costruzione e il volume vi dedica un'analisi specifica: quando i guasti sono correlati, per condivisione della rete elettrica, della connettività o dell'esposizione a eventi naturali, il vantaggio della distribuzione si riduce in misura che va stimata e non assunta.

Le ultime grandezze sono di sintesi. La metrica CFSE misura la sostenibilità dell'infrastruttura a livello di sistema, integrando efficienza energetica, intensità carbonica, consumo idrico, quota rinnovabile, recupero del calore, carbonio incorporato e utilizzazione, superando i limiti delle metriche definite sul perimetro del singolo impianto. Il punteggio confederale *CHS* articola la valutazione comparativa in tre componenti distinte, relative rispettivamente alla capacità computazionale, alle proprietà di sistema e all'idoneità ai carichi di intelligenza artificiale, perché la riduzione a un indice unico nasconderebbe proprio i compromessi che interessa rendere visibili. A queste si affianca il guadagno di federazione,

$$G_F = \text{CHS}(\mathcal{F}) / \textstyle\sum_i \text{CHS}(n_i),$$

che misura quanto la cooperazione aggiunge rispetto alla semplice somma delle parti e il cui valore, come il volume mostrerà, non è sempre maggiore dell'unità.

I.5 Struttura e articolazione del volume: le diciotto Parti e la loro logica argomentativa

L'opera si articola in diciotto Parti e sessantacinque capitoli, seguiti dalle Conclusioni e da un apparato di appendici che raccoglie i modelli formali, il quadro regolatorio e gli atlanti territoriali. La successione delle Parti non è tematica ma argomentativa: ciascuna presuppone le precedenti e prepara le successive, cosicché la lettura sequenziale ricostruisce la costruzione della tesi nell'ordine in cui essa è stata elaborata.

La prima Parte stabilisce il problema. Essa mostra perché la traiettoria di crescita della domanda computazionale incontri vincoli materiali e istituzionali che la concentrazione dimensionale non è in grado di assorbire, ricostruisce la storia dell'oscillazione fra centralizzazione e distribuzione per collocare la confederazione come quarto stadio evolutivo e separa la nozione di scala fisica da quella di scala logica fino a formulare la definizione di data center confederale.

La seconda Parte costruisce la teoria. Vi si deriva la confederazione dal quadro delle infrastrutture digitali cooperative, se ne formalizza la struttura, se ne estende alla scala

geografica lo stack architetturale, se ne studiano le leggi di scala e se ne definiscono le due grandezze portanti, l'indice di accoppiamento e il calcolo effettivamente cooperante. È la Parte più densa dal punto di vista formale ed è quella cui tutte le successive rinviano.

La terza e la quarta Parte applicano la teoria. La terza descrive analiticamente le quattro classi di nodo, dal nodo di prossimità da un centinaio di kilowatt al campus da un gigawatt e ne confronta le combinazioni a parità di potenza installata. La quarta costruisce uno scenario nazionale di riferimento, ne quantifica la capacità e lo confronta in modo sistematico con l'alternativa monolitica, mostrando in quali condizioni ciascuna configurazione prevalga.

Le Parti dalla quinta alla nona trattano le dimensioni fisiche dell'architettura. La geografia computazionale trasforma la selezione del sito nella selezione simultanea di una rete di siti, con l'apparato matematico della localizzazione. L'infrastruttura energetica federata tratta la confederazione come soggetto del sistema elettrico e non come suo semplice utente. La federazione termica affronta il raffreddamento come problema che cambia natura al variare della scala e il recupero del calore come servizio territoriale. Il tessuto di rete costruisce i bilanci di latenza e di banda che rendono la cooperazione possibile. Il calcolo e l'archiviazione confederati definiscono infine il livello che rende la confederazione utilizzabile come un'unica macchina logica, con i vincoli di località e sovranità del dato.

Le Parti dalla decima alla quattordicesima affrontano il funzionamento del sistema. L'intelligenza artificiale distribuita mappa le famiglie di tecniche sulle classi di nodo e individua le soglie oltre le quali un carico non è distribuibile. Il sistema operativo della confederazione descrive orchestrazione, decisione sotto incertezza e gemello digitale. La resilienza tratta ridondanza cooperativa, domini di guasto e continuità nativa. La sicurezza affronta fiducia nulla, consenso distribuito resiliente alla minaccia quantistica e archiviazione a privacy strutturale. La sostenibilità sistemica costruisce una metrica di sistema, il bilancio del ciclo di vita e il vincolo idrico.

Le ultime quattro Parti chiudono l'argomento. L'economia della confederazione confronta capitale e costi di esercizio fra le classi e introduce la nozione di economia di federazione accanto a quelle di scala e di scopo. La Parte dedicata a governance, territorio e società analizza gli assetti proprietari, la figura del prosumer digitale territoriale e le implicazioni di politica industriale. La Parte europea estende il modello alla confederazione delle confederazioni nazionali e ne misura il guadagno. La Parte finale sottopone la tesi a verifica, attraverso il confronto sistematico fra le configurazioni canoniche, la costruzione della mappa di dominanza e l'enunciato del teorema della scala ottimale con le sue condizioni di validità.

Chiudono il volume le appendici, che non hanno funzione ornamentale. Esse raccolgono i modelli matematici in forma estesa, la matematica dell'affidabilità, i modelli di accoppiamento, energetici, termici, di rete e di costo, il quadro regolatorio europeo e nazionale, la metodologia di simulazione, le architetture di riferimento dalla taglia del centinaio di kilowatt a quella del gigawatt e gli atlanti territoriali italiano ed europeo. Una